Local Aspects of the Global Ranking of Web Pages

Fabien Mathieu, Laurent Viennot

Local Aspects of the Global Ranking of Web Pages

Fabien Mathieu[1] and Laurent Viennot[2]

[1] France Telecom R&D
38 rue du général Leclerc, 92 130 Issy les Moulineaux France
[2] INRIA Rocquencourt
F-78153 Le Chesnay Cedex France

Abstract. Started in 1998, the search engine *Google* estimates page importance using several parameters. *PageRank* is one of those. Precisely, *PageRank* is a distribution of probability on the Web pages that depends on the Web graph. Our purpose is to show that the PageRank can be decomposed into two terms, internal and external PageRank. These two PageRanks allow a better comprehension of the PageRank signification inside and outside a site. A first application is a local algorithm to estimate the PageRank inside a site. We will also show quantitative results on the possibilities for a site to boost its own PageRank.

1 Introduction

PageRank [15] was a major algorithmic breakthrough for evaluating the importance of Web pages achieved by exploiting the topology of the Web induced by hyperlinks. Numerous works have then been devoted to better understand the relation between this Web graph structure and the quality of Web pages. Some authors have proposed alternative methods for ranking pages [12, 18] based on similar matrix computations. Other results propose different computations of an approximation of the PageRank either to obtain a faster algorithm [1, 7] or an incremental algorithm [2].

This paper tries to model how the PageRank decomposes with regards to the site partition of the Web. A site can be seen as the collection of pages on a given Web server or more generally as a set of pages tightly related. As noted by [14, 13, 16], a block structure of the Web adjacency matrix can be observed from an url-induced ordering of the pages[3], showing how an intrinsic site partition could be defined. This paper assumes that a site partition is given, ie the Web is decomposed in a collection of sites.

Even if one can naturally state that the Web graph structure is tightly related to the site partition (most of the links are local), the Web graph has mainly been studied disregarding this property. This is the case for the PageRank computation. In [1, 7], a site partition is exploited to efficiently compute an approximated PageRank. On the other hand, this paper makes an exact decomposition of the PageRank computation, showing how the PageRank can be split into an internal PageRank (related to internal links of a site) and an external PageRank (related to inter-site links). In [5, 4], the sum of the PageRanks of the pages of a site is decomposed according to internal, incoming links, outgoing links and sinks. The authors give basic hints on how the link structure of site can alter its PageRanks. A stability property of the overall PageRanks when a site changes its internal link structure is also shown. Our model of decomposed PageRank allows to push forward their analysis to better understand how a site can alter its own PageRanks.

Another contribution of our site decomposition model of PageRank is a framework for evaluating locally the global PageRank. This could be useful for a local search engine to rank the pages of a site according to a global importance knowing only locally the Web structure.

The paper is organized as follows. Section 2 defines more formally the PageRank. Section 3 introduces our model for decomposing the PageRank according to a site partition of the Web.

[3] This means clusters of pages naturally appear at different levels of the URL hierarchy, such as domain names, hosts, directories of Web servers...

Section 4 shows how to locally estimate the global PageRank of the pages of a site. Finally, Section 5 analyzes how a site administrator can alter the PageRank of its pages by modifying the links inside the site.

2 PageRank Definition

Let $G = (V, E)$ be an oriented aperiodic strongly connected graph, without self-loop, and $\mathcal{S} = (S_1, \ldots, S_k)$ a partition of G, with $k > 1$. G is supposed to be a Web graph, and $\mathcal{S}$ a site partition of G (elements of $\mathcal{S}$ are sites).

If $d^+(v)$ is the out degree of $v \in V$, we can define the following stochastic matrix $A : V \times V \to \mathbb{R}^+$

$$A = (a_{i,j})_{i,j \in V}, \text{ with } a_{(i,j)} = \begin{cases} \dfrac{1}{d^+(i)} \text{ if } i \text{ links to } j, \\ 0 \text{ otherwise.} \end{cases}$$

According to Markov processes theory [17], there is a unique probability P on V such that:

$$\forall v \in V,\ P(v) = \sum_{w \to v} \frac{P(w)}{d^+(w)} \tag{1}$$

The matrix version of this is:

$$P = A^t P, \tag{2}$$

where A^t is the transposed matrix of A.

The distribution probability P defines the PageRank of the graph G. This concept of PageRank was introduced in 1998 [15] and used by the search engine *Google* [9].

Good convergence properties and unicity of P are obtained if A is irreducible and aperiodic, that is if the underlying graph is strongly connected (and aperiodic). More inside on how A is altered to obtain such properties if G is not strongly connected is given in Section 5.2.

3 Internal PageRank, external PageRank

3.1 Notations

For $v \in V$, we call $S(v)$ the element of $\mathcal{S}$ such that $v \in S(v)$. We also define $\delta_{\mathcal{S}}$: $V \times V \to \{0, 1\}$ as follows:

$$\delta_{\mathcal{S}}(v, w) = \begin{cases} 1 \text{ if } S(v) = S(w), \\ 0 \text{ otherwise.} \end{cases}$$

Let $A_{\mathcal{S}}$ be the matrix of the projection of A on the elements of $\mathcal{S}$: $A_{\mathcal{S}} = (a_{v,w}\delta_{\mathcal{S}}(v, w))_{v,w \in V}$.

We also need to define the internal degree d_i^+ (resp. the external degree d_e^+) of a vertex v as its out degree in the graph induced by $S(v)$ (resp. $\{v\} \cup (V \setminus S(v))$).

Lastly we can define the notions of internal and external PageRank, deduced from the PageRank P seen on formula (2):

– The incoming internal PageRank P_{ii} (resp. incoming external PageRank P_{ie}) of $v \in V$ is the probability to come in v from a page of $S(v)$ (resp. $V \setminus S(v)$), that is:

$$P_{ii} = A_{\mathcal{S}}^t P \tag{3}$$

$$P_{ie} = (A - A_{\mathcal{S}})^t P = P - P_{ii} \tag{4}$$

– The outgoing internal PageRank P_{oi} (resp. outgoing external PageRank P_{oe}) of $v \in V$ is the probability to go from v to a page of $S(v)$ (resp. $V \setminus S(v)$), that is:

$$P_{oi} = (A_{\mathcal{S}}.\mathbf{1}).\times P \tag{5}$$
$$P_{oe} = ((A - A_{\mathcal{S}}).\mathbf{1}).\times P = P - P_{oi} \tag{6}$$

where $.\times$ is the element by element product.

3.2 Conservation laws

Using the definitions, we have the following equation:

$$P = P_{ie} + P_{ii} = P_{oe} + P_{oi} \tag{7}$$

We can now give the internal and external conservation laws. For each $S \in \mathcal{S}$, we see that

$$\sum_{v \in S} P(v) = \sum_{v \in S} \sum_{w \to v} \frac{P(w)}{d^+(w)} = \sum_{(w,v) \in E \cap (V \times S)} \frac{P(w)}{d^+(w)} \tag{8}$$
$$= \sum_{(w,v) \in E \cap S^2} \frac{P(w)}{d^+(w)} + \sum_{(w,v) \in E \cap (V \setminus S \times S)} \frac{P(w)}{d^+(w)} \tag{9}$$
$$= \sum_{w \in S} P_{oi}(w) + \sum_{v \in S} P_{ie}(v) \tag{10}$$

We can deduce from (7) and (10) the external conservation law:

$$\sum_{v \in S} P_{ie}(v) = \sum_{v \in S} P_{oe}(v) \tag{11}$$

We can also deduce from (7) and (11) the internal conservation law:

$$\sum_{v \in S} P_{ii}(v) = \sum_{v \in S} P_{oi}(v). \tag{12}$$

The relation (11) shows that a site gives as much PageRank (outgoing external) that it receives (incoming external). If PageRank is a random surfer flow, there is a conservation of the external PageRank flow on the graph $G/\mathcal{S}$. That remark will lead us to an intra-site and an inter-sites calculation of PageRank.

Remark: If we formalize carefully the PageRank as a flow, we have another proof of (11): the PageRank is actually a stationary flow, so the flow on every subset S is stationary, therefore we have (11). We have preferred a matrix approach for proof because matrices will be widely used in this article.

4 Local computation of the global Ranking

4.1 Relation between external PageRank and PageRank

From (3) and (4) we can write $A^t_{\mathcal{S}}.P = P - P_{ie}$, and then $P_{ie} = (Id - A^t_{\mathcal{S}})P$, where Id is the identity matrix.

Lemma 1. *The matrix $(Id - A^t_{\mathcal{S}})$ is inversible.*

Proof. As G is strongly connected, there are links between sites. Thus we have $0 < A_{\mathcal{S}} < A$. $A_{\mathcal{S}}$ is strictly sub-stochastic, so its spectral radius is strictly inferior to 1. Therefore $(Id - A_{\mathcal{S}}^t)^{-1}$ exists.

Lemma 1 allows us to express P as a function of P_{ie}:

$$P = (Id - A_{\mathcal{S}}^t)^{-1} P_{ie} \tag{13}$$

Knowing the incoming external PageRank P_{ie} of a site S, we can theoretically compute the PageRank of the pages of S with only the local graph G_S.

Remark $(Id - A_{\mathcal{S}}^t)^{-1} = \sum_{k=0}^{\infty} (A_{\mathcal{S}}^t)^k$ is a diagonal by blocks matrix, that can be interpreted as the transition matrix of all the internal paths.

4.2 External PageRank matrix

We want to translate the intuition of Equation (11) in a conservation law with P_{ie} only. From (4) and (13), we have:

$$P_{ie} = (A - A_{\mathcal{S}})^t P = (A - A_{\mathcal{S}})^t (Id - A_{\mathcal{S}}^t)^{-1} P_{ie} \tag{14}$$

We thus define the external PageRank transition matrix A_e:

$$A_e^t = (A - A_{\mathcal{S}})^t (Id - A_{\mathcal{S}}^t)^{-1}$$

Lemma 2. *The matrix A_e is stochastic.*

Proof. We just have to show that the sum of each column of A_e^t is 1. First, we rewrite A_e^t:

$$\begin{aligned} A_e^t &= \sum_{k=0}^{\infty} \left(A^t (A_{\mathcal{S}}^t)^k - (A_{\mathcal{S}}^t)^{k+1} \right) \\ &= A^t + \sum_{k=1}^{\infty} \left(A^t (A_{\mathcal{S}}^t)^k - (A_{\mathcal{S}}^t)^k \right) \\ &= A^t + A^t M - M, \text{ with } M = \sum_{k=1}^{\infty} (A_{\mathcal{S}}^t)^k \end{aligned}$$

Then we consider the sum s_w of a column w in $A^t M$:

$$s_w = \sum_{u \in V} \sum_{v \in V} A_{u,v}^t M_{v,w} = \sum_{v \in V} \left(\sum_{u \in V} A_{u,v}^t \right) M_{v,w} = \sum_{v \in V} M_{v,w}$$

So the sum of each column of $A^t M - M$ is null; then A_e^t is stochastic as A^t.

4.3 Partially distributed PageRank algorithm

From (13) and (14) we can suggest a half-distributed algorithm for computing the PageRank:

- Each site S computes from its block of $A_{\mathcal{S}}$ a block of the matrix $(Id - A_{\mathcal{S}}^t)^{-1}$.
- The coefficients of A_e are centralized.
- The external PageRank P'_e associated with A_e is centrally computed using $A_e^t P'_e = P'_e$.
- Each site S gets its own PageRank thanks to the relation $P' = (Id - A_{\mathcal{S}}^t)^{-1} P'_e$ applied to its block.

Lemma 3. *The vector P' we obtain is, once normalized, the PageRank P of G.*

Proof. We have to show that P' is an eigenvector of A^t, and that its eigenvalue is 1:

$$\begin{aligned} A^t P' &= A^t (Id - A_{\mathcal{S}}^t)^{-1} P'_e \\ &= (A^t - A_{\mathcal{S}})(Id - A_{\mathcal{S}}^t)^{-1} P'_e + A_{\mathcal{S}}^t (Id - A_{\mathcal{S}}^t)^{-1} P'_e \\ &= A_e^t P'_e + \left((Id - A_{\mathcal{S}}^t)^{-1} - (Id - A_{\mathcal{S}}^t)(Id - A_{\mathcal{S}}^t)^{-1} \right) P'_e \\ &= P'_e + \left((Id - A_{\mathcal{S}}^t)^{-1} - Id \right) P'_e \\ &= P'_e + P' - P'_e = P' \end{aligned}$$

As the principal eigenvalue of A, that is 1, is unique, P and P' are homothetic, so $P = P'$ (after normalization).

Efficient computation of external PageRank: The above algorithm could seem useless since the centralized step operates on A_e which is a $|V| \times |V|$ matrix. However, from $P'_e = A_e^t P'_e$, we can write $P'_e = (A - A_{\mathcal{S}})^t P''$, where $P'' = (Id - A_{\mathcal{S}}^t)^{-1} P'_e$. That means that only vertices with incoming external links have a non null value in P'_e. Thus we can safely reduce A_e and P'_e to their projections on V_{ext}, where V_{ext} is the set of vertices that have at least one incoming external link. The new equation to compute the external PageRank is then:

$$P_{Reduced} = A_{Reduced}^t P_{Reduced}$$

where $P_{Reduced}$ is a vector of V_{ext} and $A_{Reduced}$ is a $V_{ext} \times V_{ext}$ matrix defined by $A_{Reduced}(i,j) = A_e(i,j)$ for each $i, j \in V_{ext}$. Analysis of Web logs and crawls shows that for the site inria.fr we have $|V_{ext}| \leq 0.1\,|V|$.

4.4 Link between our algorithm and BlockRank algorithms

Kamvar et al. [1], and more recently *Broder et al.* [7], use a similar algorithm based on the local block structure of the Web [13]. They first compute a local PageRank, that with our notation is the PageRank of the $A_{\mathcal{S}}$ matrix, then use it to compute a BlockRank matrix B that is a sub-stochastic matrix on the quotient graph $G/\mathcal{S}$ defined as transitions between sites.. Their approximation of PageRank is the local PageRank weighted by the probability of being in a given block $S \in \mathcal{S}$, obtained using B (all details can be found in [1, 7]). Although they look similar, their algorithms and ours have some differences that we would like to point out:

- The algorithm presented in Section 4.3 converges to the exact PageRank. Thus the problem of the weaker PageRank on root pages in BlockRank [1, 7] does not occur.
- The local PageRank matrix $A_{\mathcal{S}}$ and the BlockRank matrix B used in BlockRank correspond respectively to our internal transition matrix $(Id - A_{\mathcal{S}}^t)^{-1}$ and our external PageRank matrix A_e. This difference comes from our modeling of the PageRank as a flow.

***À la BlockRank* algorithm:** we just saw that B is a $k \times k$ matrix, when $A_{Reduced}$ is a $V_{ext} \times V_{ext}$ matrix, with $k \leq |V_{ext}| \leq |V|$. This size difference is the price of the exact computation of PageRank. Indeed $A_{Reduced}$ can still be reduced up to a $k \times k$ matrix if we centralise all the external flow to one or a few chosen page(s) per site (the main page(s) for example). Reduction is done by choosing a set of main pages V_{main}, with $m(S) \geq 1$ page(s) per site, and using the projection A_{main} of the matrix $A_e.R$ to V_{main}, where:

$$R(v,w) = \begin{cases} \dfrac{1}{m(S(w))} \text{ if } S(v) = S(w) \text{ and } w \in V_{main} \\ 0 \text{ otherwise} \end{cases}$$

The size of the external matrix becomes $|V_{main}| \times |V_{main}|$, with V_{main} fully configurable down to one page per site (thus reaching the $k \times k$ limit), but it is now an approximation, that should lead to an overestimation of the PageRank of the main pages (when BlockRank underestimates them). Our formalism thus allows to obtain an approximated PageRank algorithm similar to BlockRank with additional flexibility from fast computation time to exact PageRank computation.

4.5 Estimation of a site PageRank

A natural question is to ask if a site S can estimate the ranking of its pages only knowing local data. This can be very valuable for an internal search engine to be able to estimate the global ranking of its pages without crawling all the Web or asking an external search engine. From (13), all we need is an estimation of the incoming external PageRank.

According to [15], PageRank models the statistic behaviour of surfers crawling the Web. It seems then natural to estimate the PageRank of a page by the average hits it gets. More specifically, the incoming external PageRank should be proportional to the average hits from outside the site. So each site can get an estimation of the incoming external rank from analysing the logs files of its Web server. This gives webmasters a smart way for locally computing a structural ranking flavoured with real traffic statistics.

Abiteboul et al. [2] states that the incoming degree is a good estimation of the PageRank. Thus the number of external references for each page (obtained from the logs files) is another estimation of P_{ie}. Compared to previous estimation, incoming degree estimation is purely structural.

Both estimation methods of the incoming external PageRank will be furthered studied in future works.

Once P_{ie} is estimated, the global PageRank can be obtained by $P = (Id - A_{\mathcal{S}}^t)^{-1} P_{ie}$. In fact, $(Id - A_{\mathcal{S}}^t)^{-1}$ does not have to be calculated explicitly. It is better to solve $P = A_{\mathcal{S}}^t P + P_{ie}$ using iterative methods, for example by choosing P_0 and iterating

$$P_{n+1} = A_{\mathcal{S}}^t P_n + P_{ie}$$

This converges, because the spectral radius of $A_{\mathcal{S}}$ is strictly inferior to 1. Empirical results from [15] suggest a fast convergence of that sort of algorithm applied to portions of the Web graph.

Remark There are lot of methods to improve the convergence of that sort of iterative computation [3, 11]. The purpose of this paper is not to optimize this part of the computation, so we only give the basic Jacobi method.

Interest of our method One could wonder why not keeping the average hits per page as an estimation of the PageRank? We believe our method can give a better PageRank to pages newly created, that do not get a lot of hits yet but are well linked and will surely get known.

Another advantage is that the P_{ie} input can be very flexible. The webmaster could manually alter P_{ie} to promote some pages while keeping a minimum of ranking.

5 Locally altering the PageRank

Our decomposition of the PageRank explains some results about the ability that a site has to alter its own PageRank. A first approximation is to say that if a site can hardly alter the external PageRank, this is much easier for the internal PageRank.

5.1 Amplification factor

Let S be a site, $P(S) = \sum_{v \in S} P(v)$ and $P_{ie}(S) = \sum_{v \in S} P_{ie}(v)$. We can define the amplification factor of S by $\alpha(S) = \frac{P(S)}{P_{ie}(S)}$. This factor depends on both S and the distribution of the actual external PageRank[4], but knowing S we can estimate $\alpha(S)$.

Lemma 4. *The amplification factor can be estimated by:*

$$\frac{1}{1-\omega} \leq \alpha(S) \leq \frac{1}{1-\Omega} \tag{15}$$

with $\omega = \min_{v \in S} \frac{d_i^+(v)}{d^+(v)}$ *and* $\Omega = \max_{v \in S} \frac{d_i^+(v)}{d^+(v)}$.

Proof. If we see $\mathcal{S}$ as a $|\mathcal{S}|$-dimensional vector space, for each base vector e_v, $v \in S$, we have $\|A_{\mathcal{S}}(e_v)\|_1 = \frac{d_i^+(v)}{d^+(v)}$, therefore $\omega \|X\|_1 \leq \|A_{\mathcal{S}} X\|_1 \leq \Omega \|X\|_1$ for any vector $X > 0$ defined on S.

The first inequality of (15) is obtained as follows:

$$P(S) = \sum_{v \in S} P(v) = \left\| \sum_{k \in \mathbb{N}} (A_S^t)^k (P_{ie}) \right\|_1 \geq \sum_{k=0}^{\infty} \omega^k \|P_{ie}\|_1 = \frac{1}{1-\omega} P_{ie}(S)$$

The proof of the second inequality is similar.

The consequences of this amplification system is that a site can arbitrarily increase its PageRank. In the limit case where the site has no external link[5], we have a short-circuit phenomenon. This is known as the sink hole phenomenon [15]: a set of pages with no outgoing link absorbs all PageRank.

Fortunately, we will see how the damping factor reduces this effect.

5.2 The damping factor

As said in Section 2, good convergence properties are obtained whenever G is a strongly connected graph. Otherwise, transient page can exist (they obtain a zero PageRank), and A may be sub-stochastic (if some pages do not have out-link). As the Web graph is far from being strongly connected [8], there are several techniques to overcome this, often by altering the transition matrix A. We focus on the damping factor[6], introduced by [6]. It is originally used by *Google* on a graph where leaves are non-recursively removed and reinjected after P converged. The principle of the damping factor is to replace A by $d.A + \frac{1-d}{|V|}\mathbf{1}.\mathbf{1}^t$, where $\mathbf{1}$ is a vector filled with ones and d the damping factor. The new transition matrix represents a weighted strongly connected graph, and it is stochastic (we still suppose that A is stochastic; see [15] for normalization issues about pages without outgoing links). We have then a superposition of classic transitions ($d.A$) and damping transitions ($\frac{1-d}{|V|}\mathbf{1}\mathbf{1}^t$). Damping transitions are supposed to model the action of moving anywhere in the Web without following

[4] Note that the distribution of the external PageRank can more or less depend on the distribution of S if S is close to pages pointing to it.

[5] A real site does not have to respect the strong connectivity of A. In particular, many commercial sites do not have any external link [8].

[6] For a non-exhaustive view of the other techniques:

- *Page et al.* [15] suggests to compensate the flow leak in A by normalizing P at each iteration.
- *Haveliwala et al.* [10] turn A explicitly into a stochastic matrix by removing recursively pages without link.
- *Abiteboul et al.* [2] adds a virtual damping page that links to and is linked to every other page.

any static link (user *Bookmarks*, search engines, keyboard input,...). Note that $(\frac{1}{|V|}\mathbf{1}\mathbf{1}^t)$ corresponds to the uniform transition matrix from any page to any page.

Instead of splitting the damping flow into an external one and an internal one, we find more interesting to introduce the notions of induced PageRank P_{ind} and dissipated PageRank P_{dis}. We have now six different PageRanks corresponding to three types of flows as shown in Figure 1:

flow	incoming	outgoing
internal	$P_{ii} = dA_S^t P$	$P_{oi} = d(A_S\mathbf{1}).\times P$
external	$P_{ie} = d(A - A_S)^t P$	$P_{oe} = d((A - A_S)\mathbf{1}).\times P$
damping	$P_{ind} = \frac{1-d}{\|V\|}\mathbf{1}$	$P_{dis} = (1-d)P$

Fig. 1. The different flows of PageRank in the damping factor case

We can rewrite the conservation laws considering the whole damping flow as external. Of course there are internal damping transitions, but we choose to tag them as external. Thus the internal flow conservation law does not change, but we have a new external flow conservation law:

$$\sum_{v\in S}\left(P_{ie}(v) + P_{ind}(v)\right) = \sum_{v\in S}\left(P_{oe}(v) + P_{dis}(v)\right),$$

that we will note

$$P_{ie}(S) + P_{ind}(S) = P_{oe}(S) + P_{dis}(S) \tag{16}$$

PageRank stability The equation (16) shows the stability of the classic flow at the site level. From $P_{ind}(S) = (1-d)\frac{|S|}{|V|}$ and $P_{dis}(S) = (1-d)P(S)$, we can tell that for a site whose PageRank $P(S)$ is above (resp. below) the average PageRank (which is $\frac{|S|}{|V|}$ for a site of size $|S|$), the outgoing external PageRank $P_{oe}(S)$ is inferior (resp. superior) to the incoming external PageRank $P_{ie}(S)$. In other words, a *rich* site (in term of PageRank) will be greedy and will give less than it receives (damping excluded), and vice versa. The damping factor causes a retro-action that limits the phenomenon of over-amplification, as developed in next section.

5.3 Damping and amplification

The transition matrix is of the form $d.A + \frac{1-d}{|V|}\mathbf{1}\mathbf{1}^t$. We obtain results similar to Section 5.1 by replacing A by dA and P_{ie} by the total incoming external PageRank $P_{ie} + P_{ind}$.

Lemma 5. *The amplification factor* $\alpha'(S) = \frac{P(S)}{P_{ie}(S)+P_{ind}(S)}$ *verifies:*

$$\frac{1}{1-d\omega} \leq \alpha'(S) \leq \frac{1}{1-d\Omega}. \tag{17}$$

Proof. It is the same that for (15); we can write:

$$P(S) = \sum_{v \in S} P(v) = \left\| \sum_{k \in \mathbb{N}} (dA_S^t)^k (P_{ie} + \frac{1-d}{|V|}\mathbf{1}) \right\|_1$$
$$\leq \sum_{k=0}^{\infty} (d\Omega)^k (\|P_{ie}\|_1 + (1-d)\frac{\|\mathbf{1}\|_1}{|V|})$$
$$\leq \frac{1}{1-d\Omega} \left(P_{ie}(S) + (1-d)\frac{|S|}{|V|} \right)$$

The second inequality is obtained similarly.

Numerical Value It is not impossible for a real site to have $\omega = \Omega = 0$ (site without internal link) or $\omega = \Omega = 1$ (site without external link). So the amplification factor can vary between 1 and $\frac{1}{1-d}$. The empirical value of d being 0.85, we deduce that with a fixed incoming external PageRank, the PageRank of a site can fluctuate up to a factor $\frac{20}{3}\ldots$

PageRank robustness *Bianchini et al.* [4] states that the effect that a site can produce onto the Web is bounded by the PageRank of this site. If we consider two instants t and $t+1$, they show that:

$$\sum_{v \in V} |P_t(v) - P_{t+1}(v)| \leq \frac{2d}{1-d} \sum_{s \in S} P_t(s)$$

This result is a straightforward implication of Lemma 5: if the site S changes between t and $t+1$, the PageRank variation inside S is at most $\frac{d}{1-d}P(S)$, implying a variation up to another $\frac{d}{1-d}P(S)$ outside the site, since the total PageRank stays equal to 1.

5.4 Amplification of a given page

However, a search engine answers a lot of pages for most of the requests. This implies that a site administrator may be less interested by getting a large average PageRank than getting a few pages with high PageRank or even a single one. We thus consider the following problem: let S be a site of $n+1$ pages and P_{ie} its incoming external PageRank; how can we maximize the PageRank of a given page $v_0 \in S$?

The answer is not difficult once we remark the optimal link structure is when v_0 links to all other pages of S and all other pages of S link to v_0 and only v_0[7]. It is not hard then to have a limitation of $P(v_0)$:

$$P(v_0) \leq \frac{P_{ie}(S)}{1-d^2} + \frac{1+nd}{(1+d)\,|V|}, \tag{18}$$

with equality if and only if $P_{ie}(S) = P_{ie}(v_0)$.

This suggests some strategies to improve the PageRank of a page v_0[8]. For instance:

- If v_0 links to all other pages without backward links[9], adding the links to v_0 can increase the PageRank of v_0 up to $\frac{1}{1-d^2} \simeq 3,6$.
- The optimal strategy ensures for v_0 a minimal PageRank at least equal to the average PageRank $\frac{1}{|V|}$ even if P_{ie} is null.
- If $1 \ll n \leq |V|$ (for a large site dynamically generating pages linking to v_0), the ratio $\frac{P(v_0)}{P_{average}}$ is about $\frac{d}{1+d}n$.

[7] PageRank algorithms systematically remove self-loops, so a single page cannot amplify itself.

[8] In fact, it seems that *Google* is rather aware of these strategies, so they do not work as well as they should in theory...

[9] A typical situation when using *frames*.

6 Conclusion

We have proposed a decomposition of the PageRank flow in accordance with the notion of site, showing how to use it for estimating locally the global PageRanks inside a site. However, this relies on estimating the incoming PageRank either with real user hits or external referer counts. Further experiments are needed for fully validating this approach. Another interesting research direction includes distributed computation of the PageRank: assuming that several sites collaborate, how to compute the PageRank induced by their union? Our model is certainly the first step for that. It can also be useful for evaluating approaches that alter the PageRank computation based on a site decomposition as proposed by [1, 7] for speeding up the computation. Another related issue is the identification and the ranking of sites rather than pages.

At least, the flow decomposition has allowed to analyze some strategies that the webmasters could use if an unrefined version of PageRank was used by search engines. We have shown that the PageRank defined in [15] can be very versatile when subject to non-cooperative strategies. It also seems that P_{ie} can be more robust, assuming we are able to find a site partition $\mathcal{S}$ that reflects the reality.

References

1. Technical report.
2. S. Abiteboul, M. Preda, and G. Cobena. Adaptive on-line page importance computation. In *Proc. 12th International World Wide Web conference*, 2003.
3. G. Allaire and S. M. Kaber. *Algèbre linéaire numérique*. Ellipses, 1998.
4. M. Bianchini, M. Gori, and F. Scarselli. Pagerank: A circuital analysis. In *Proc. 11th International Word Wide Web Conference*, 2002.
5. M. Bianchini, M. Gori, and F. Scarselli. Inside pagerank. volume 5, pages 92–128, New York, NY, USA, 2005. ACM Press.
6. S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.
7. A. Z. Broder, R. Lempel, F. Maghoul, and J. Pedersen. Efficient pagerank approximation via graph aggregation. In *Proc. 13th International World Wide Web conference on Alternate track papers & posters*, pages 484–485, 2004.
8. A. B. et al. Graph structure in the web. In *Proc. 9th International World Wide Web Conference*, pages 309–320, 2000.
9. Google. http://www.google.com/, 1998.
10. T. Haveliwala. Efficient computation of pageRank. Technical Report 1999-31, Stanford University, 1999.
11. S. Kamvar, T. Haveliwala, C. Manning, and G. Golub. Extrapolation methods for accelerating pagerank computations. In *Proc. 12th International World Wide Web Conference*, pages 261–270, 2003.
12. J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, San Francisco, California, 25–27 Jan. 1998.
13. F. Mathieu and L. Viennot. Structure intrinsèque du web. Technical Report RR-4663, INRIA, Dec. 2002.
14. F. Mathieu and L. Viennot. Local structure in the web. In *Proc. 12th International World Wide Web conference on Alternate track papers & posters*, 2003. poster.
15. L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Computer Science Department, Stanford University, 1999.
16. S. Raghavan and H. Garcia-Molina. Representing web graphs, 2003.
17. L. Saloff-Coste. Lectures on finite Markov chains. In G. G. E. Giné and L. Saloff-Coste, editors, *Lecture Notes on Probability Theory and Statistics*, number 1665 in LNM, pages 301–413. Springer Verlag, 1996.
18. P. Senellart and V. D. Blondel. Automatic discovery of similar words. In Michael W. Berry, editor, *Survey of Text Mining*. Springer-Verlag, Aug. 2003.